

Statistical mechanics of random geometric graphs: Geometry-induced first order phase transition

Massimo Ostilli^{1,2} and Ginestra Bianconi³

¹ *Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis 88040-900, Santa Catarina, Brazil*

² *Dipartimento di Fisica, Università di Roma La Sapienza, Piazzale Aldo Moro 2, Roma I-00185, Italy*

³ *Queen Mary University of London, London, E1 4NS, United Kingdom*

Random geometric graphs (RGG) can be formalized as hidden-variables models where the hidden variables are the coordinates of the nodes. Here we develop a general approach to extract the typical configurations of a generic hidden-variables model and apply the resulting equations to RGG. For any RGG, defined through a rigid or a soft geometric rule, the method reduces to a non trivial satisfaction problem: Given N nodes, a domain \mathcal{D} , and a desired average connectivity $\langle k \rangle$, find - if any - the distribution of nodes having support in \mathcal{D} and average connectivity $\langle k \rangle$. We find out that, in the thermodynamic limit, nodes are either uniformly distributed or highly condensed in a small region, the two regimes being separated by a first order phase transition characterized by a $O(N)$ jump of $\langle k \rangle$. Other intermediate values of $\langle k \rangle$ correspond to very rare graph realizations. The phase transition is observed as a function of a parameter $a \in [0, 1]$ that tunes the underlying geometry. In particular, $a = 1$ indicates a rigid geometry where only close nodes are connected, while $a = 0$ indicates a rigid anti-geometry where only distant nodes are connected. Consistently, when $a = 1/2$ there is no geometry and no phase transition. After discussing the numerical analysis, we provide a combinatorial argument to fully explain the mechanism inducing this phase transition and recognize it as an easy-hard-easy transition. Our result shows that, in general, ad hoc optimized networks can hardly be designed, unless to rely to specific heterogeneous constructions, not necessarily scale free.

PACS numbers: 89.75.Hc, 05.70.Fh, 05.70.Ce, 02.70.-c

I. INTRODUCTION

In the last decades, statistical mechanics has seen an outstanding progress as a powerful mathematical tool, physically oriented, for analyzing complex systems of the most diverse nature. A special impulse was later given to the interplay between concepts close to the realm of statistical physics, mostly within the spin glass theory [1], and the mathematical theory of complexity for optimization problems [2, 3]. In more recent years, the developments in complex network theory and its numerous applications have made even stronger the importance of statistical mechanics as a very interdisciplinary mathematical tool [4–9]. In network theory, elements and interactions are simplified as nodes and links between them. Several models for growing or static networks have been proposed and turned out able to reproduce the main universal features observed in real networks. One of the most promising frameworks to analyze networks is provided by hidden variables models [10–13]. The general idea behind hidden variable models relies on the possibility that each node is attributed a value h that expresses its propensity to be connected to another node with hidden variable h' via a formula that relates h with h' . It turns out that this scheme is both very general and very powerful in the description of many diverse networks, especially with respect to the possibility of having analytic results.

In this paper, we use statistical mechanics to face and solve two concatenated problems in hidden variable models. We first develop a general approach to extrapolate the typical configurations of a generic hidden-variables model, *i.e.*, we use statistical mechanics to derive, in the

thermodynamic limit, the equations that describe, in the space of the hidden variables h , the node distribution ν_h corresponding to the most important graph realizations (the others being exponentially rare in the thermodynamic limit). Then, we apply these general equations to random geometric graphs (RGG) [14–16] where the h 's represent vector positions. Unlike random or common complex network models, in RGG nodes are connected according to an underlying geometry (Euclidean or not). In particular, in classical RGG two nodes are connected if their distance is at most equal to a given threshold. RGG are important in many theoretical and practical aspects, but despite that, the exact analysis of RGG has been mostly confined to the Euclidean $d = 1$ and $d = 2$ dimensional cases, and mainly focused on the percolation problem. Notice that, when $d > 1$, RGG are not tree-like networks. In fact, they have dense loops, *i.e.*, the probability that a random walk of l links is closed remains finite for any $l \geq 3$, even in the infinite size limit (for $l = 3$ this amounts to a finite clustering coefficient). This fact avoids using the same probabilistic techniques, like the Belief-Propagation (or Bethe-Peierls) methods, which are exact only on tree-like graphs. It is worth mentioning here the Kasteleyn and Fortuin approach to the percolation problem [17], which maps the counting of connected clusters toward the mean-field Potts model; an outstanding example of how statistical mechanics can be used to solve combinatorial problems (see also [2]). However, when we are dealing with RGG, due to the presence of dense loops, the Kasteleyn and Fortuin approach would require to solve a non mean-field Potts model, an analytically unfeasible task.

In this paper we adopt a different approach. We do not aim to analyzing the percolation problem in RGG, nor to have a full descriptive solution of RGG (which would allow us to know all the graph metrics). Rather, we look for reduced descriptive, yet exact, solutions. By using the hidden variable framework, and our general result for the typical configurations, we face a practical satisfaction problem: Given N nodes, a domain \mathcal{D} , and a desired average connectivity $\langle k \rangle$, find - if any - the distribution of nodes having support in \mathcal{D} and average connectivity $\langle k \rangle$. Notice that this is a typical problem in ad hoc wireless optimized networks [18]: Embedded in a space with $d = 2$ or $d = 3$ dimensions, one has N mobile devices that must be connected, in average, with a certain number of neighbors at a minimal cost. Furthermore, it might also be required that the whole structure be a connected graph. However, we find out that, in the thermodynamic limit, nodes can only be either uniformly distributed or highly condensed in a small region, the two regimes being separated by a first order phase transition characterized by a $O(N)$ jump of $\langle k \rangle$ (see Fig.1 for an example whose details will be discussed later). Other intermediate values of $\langle k \rangle$ correspond only to very rare graph realizations. The phase transition is triggered by the underlying geometry whose strength can be tuned by a parameter $a \in [0, 1]$, with $a = 1$ for rigid geometry (only close nodes are connected) and $a = 0$ for rigid anti-geometry (only distant nodes are connected). Consistently, when $a = 1/2$ there is no geometry and no phase transition. After discussing the numerical solution for RGG in $d = 1$ dimensions, we provide a combinatorial argument to fully explain the mechanism inducing this phase transition in general RGG. In turn, the combinatorial argument leads us to recognize the phase transition as an easy-hard-easy transition, the hard solutions of the problem being located in between the uniform and condensed regimes, *i.e.*, at the critical point. Furthermore, we show that the hard solutions corresponding to connected graphs are networks quite heterogeneous. Our result shows that, in general, ad hoc optimized networks can hardly be designed, unless to rely to strongly heterogeneous constructions, not necessarily power law.

II. HIDDEN-VARIABLES MODELS

Given N nodes, hidden variable models are defined in the following way: *i)* to each node we associate a hidden variable h drawn from a given probability density function (PDF) $\rho(h)$; *ii)* between any pair of nodes, we assign a link according to a given set of probabilities $p(h, h')$, where h and h' are the hidden variables associated to the two nodes. The probabilities $p(h, h')$ can be any function of the h 's, the only requirement being that $0 \leq p(h, h') \leq 1$. The hidden variables can be real numbers or also vectors. Many networks can be embedded in a hidden-variables scheme. For example, in RGG the h 's are vectors representing the positions of the nodes,

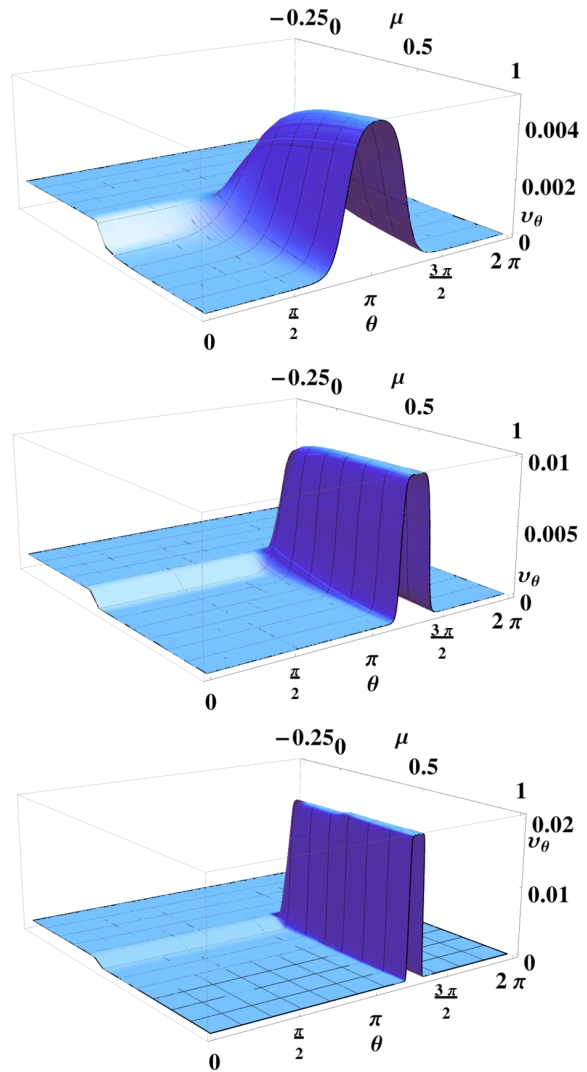


FIG. 1: (Color online) Typical node-distribution (associated to the typical graphs) for a RGG defined over the circle according to a simple geometric rule (44)-(45) with a rigid constrain ($a = 1$). The plot gives the distribution ν_θ as a function of the angle θ for each value of the chemical potential μ . As the number of nodes N increases, there emerges a separation of two phases in correspondence of a critical value of the chemical potential $\mu_c \simeq 0.25$. For $\mu < \mu_c$ the nodes tend to be uniformly distributed and have a minimal average connectivity $\langle k \rangle = O(1)$, whereas, for $\mu > \mu_c$, a condensation of nodes takes place and tend to form a fully connected graph with $\langle k \rangle = O(N)$. Top to bottom: $N = 10, 20, 40$.

and the $p(h, h')$'s are defined in terms of geometric rules (non deterministic if $0 < p(h, h') < 1$, or deterministic if $p(h, h')$ takes only the values 0 or 1). Notice that, by construction, the steps *i)*-*ii)* generate only simple graphs (*i.e.*, without multiple edges or loop-edges).

Particular attention has been paid to the “configuration model”. In this case $p(h, h')$ has the Fermi-Dirac

form (or similar generalizations)

$$p(h, h') = \left(1 + \frac{k_s^2}{hh'}\right)^{-1}, \quad k_s = \sqrt{N\bar{k}}, \quad (1)$$

where, for large N , \bar{k} coincides with the average degree. In general, the actual degree k of the nodes of the network realized with the above scheme are distributed according to ρ and average degree equal to \bar{k} . In particular, if we choose the following PDF having support in $[h_{\min}, h_{\max}]$

$$\rho(h) = a h^{-\gamma}, \quad h_{\max} \geq h \geq h_{\min} > 0, \quad (2)$$

with $\gamma > 2$, the degree-distribution $p(k)$ of the resulting network will be a power law with exponent γ .

For what follows, is important to note that Eqs. (1) and (2) represent a very particular hidden-variables model. In fact, for such a specific example we have full control of the average number of links via the parameter \bar{k} inside the function $p(h, h')$. Notice that, given any couple $p(h, h')$ and $\rho(h)$, the hidden variable scheme defined through the steps *i-ii*) determines an ensemble of graphs whose properties, like average connectivity, degree distribution, density of triangles etc..., are all mathematically encoded in $p(h, h')$ and $\rho(h)$. However, only in a few cases the relation between these ensemble properties and the parameters entering the functions $p(h, h')$ and $\rho(h)$ turn out to be relatively simple. In general, given $p(h, h')$ and $\rho(h)$, we are not able to control directly the ensemble properties via these parameters, not even the average number of links. In fact, this is the case for RGG. Therefore, for the sake of generality, we will allow for the presence of a chemical potential μ to be adjusted toward a suitable value in order to have a desired average number of links.

III. TYPICAL CONFIGURATIONS IN HIDDEN-VARIABLES MODELS

A. Motivation, formulation of the problem and synthesis of the method

Let be given a hidden-variables model through the functions $p(h, h')$ and $\rho(h)$. There are two kinds of network ensembles that can be generated. Inspired by statistical mechanics, we call these two ensembles: canonical and gran canonical.

Canonical Ensemble. If we make a *una tantum* sampling of the h 's and these are kept fixed, we are dealing with an ensemble of graphs where the h 's play the role of parameters, while the random variables are the elements of the adjacency matrix $\{a_{i,j}\}$, each taking the values 1 or 0 for the presence or absence of a link, respectively. Examples include networks whose nodes have fixed expected degrees and, more in general, networks that can be described as exponential random graphs [23]. One peculiar feature of the Canonical Ensemble is that the relative fluctuations of graph-observables that are extensive in the system size tend to 0 for $N \rightarrow \infty$. This fact allows to extrapolate many characteristics of the ensemble which we can define as “typical” [5, 7].

Gran Canonical Ensemble. This is the kind of ensemble we will deal with. In this case the h 's are random variables of the ensemble together with the $\{a_{i,j}\}$. An important example are RGG where the h 's represent the coordinates of the nodes. Another example is provided by gradient networks, where the h 's determine the flow directions [19]. However, the Gran Canonical Ensemble is not just an enlargement of the Canonical one; it has a quite different nature. Unlike the Canonical Ensemble, in the Gran Canonical Ensemble, in general, fluctuations of extensive quantities cannot be neglected. In particular, it has been shown that, in the scale-free model (1)-(2), the number of links and the entropy are the only observables whose relative fluctuations can be neglected when $N \rightarrow \infty$ [20–22]. In order to calculate the ensemble averages, $\langle \cdot \rangle$, one has to evaluate a generalized partition function, or functional generator, $Z(\{x_{i,j}\}; \{y_i\})$, where the $x_{i,j}$ and the y_i are external fields conjugated to the variables $a_{i,j}$ and h_i , respectively. A manifestation of the large fluctuations of the Gran Canonical Ensemble is that $Z(\{x_{i,j}\}; \{y_i\})$ cannot be calculated via saddle point techniques (in Appendix A we show this in detail). However, under certain conditions, we can still find some typical features of the ensemble, exactly. In fact, if we limit the description of a graph to its hidden-variables sequence, such a plan is mathematically possible. To this aim, we analyze the generating function $Z(\mu)$ where μ is conjugated to the number of links $\sum_{i<j} a_{i,j}$ and plays the role of the chemical potential of the ensemble. Notice that, in general, there is a one-to-one correspondence between μ and the average connectivity $\langle k \rangle$: $\mu = \mu(\langle k \rangle)$. We then perform the change of variables $\{h_i\} \rightarrow \{N_h\}$, where $N_h = \sum_i \delta(h_i, h)$. With this change

of variables, for any $\mu \in I$, where I is a finite interval, the integrand in $Z(\mu)$ is highly concentrated around its maximum, and $Z(\mu) \propto \exp(-F(\{N_h^*\}))$, where $F(\{N_h^*\})$ plays the role of a free energy with respect to the variables $\{N_h\}$, and the $N_h^* = N\nu_h^*$ are solution of certain saddle point Eqs.. As we will show, given N , $\mathbf{p} = \{p_{h,h'}\}$, $\boldsymbol{\rho} = \{\rho_h\}$, and a desired average connectivity $\langle k \rangle$, the saddle point Eqs. for ν_h and μ lead to a crucial duality relation between the ensemble characterized by $\mathbf{p} = \{p_{h,h'}\}$, $\boldsymbol{\rho} = \{\rho_h\}$ and $\mu = \mu(\langle k \rangle)$, and an ensemble characterized by $\mathbf{q}_\mu = \{q_{h,h'}\}$, $\boldsymbol{\nu}_\mu = \{\nu_h^*\}$, and $\mu = 0$:

$$\text{Ensemble } \{\mathbf{p}, \boldsymbol{\rho}, \mu\} = \text{Ensemble } \{\mathbf{q}_\mu, \boldsymbol{\nu}_\mu, 0\}, \quad (3)$$

where \mathbf{q}_μ are a proper modification of the probabilities \mathbf{p} . To simplify the notation we shall often omit the suffix μ in $\boldsymbol{\nu}_\mu$ and \mathbf{q}_μ . The duality relation (3) allows to draw, in principle, networks with arbitrary $\langle k \rangle$ without changing N . In particular, as we shall see, for rigid RGG we have $\mathbf{q}_\mu = \mathbf{p}$, but $\boldsymbol{\nu}_\mu \neq \boldsymbol{\rho}$. In the following of this Section we derive the saddle point Eqs. that lead to Eq. (3) and the relation between $\langle k \rangle$ and μ .

B. Analysis

Let be given a hidden-variables model through the functions $p(h, h')$ and $\rho(h)$. The conditional probability to realize a graph $\mathbf{G} = \{a_{i,j}\}$ with frozen hidden variables $\{h_i\}$ is

$$P(\mathbf{G}|\{h_i\}) = \prod_{i < j} [p(h_i, h_j)]^{a_{i,j}} [1 - p(h_i, h_j)]^{1-a_{i,j}} \quad (4)$$

while the (unconditioned) probability to have \mathbf{G} is

$$P(\mathbf{G}) = \int \prod_i dh_i \rho(h_i) \prod_{i < j} [p(h_i, h_j)]^{a_{i,j}} \times [1 - p(h_i, h_j)]^{1-a_{i,j}}. \quad (5)$$

The generating function we want to calculate is

$$Z(\mu) = \int \prod_i dh_i \rho(h_i) \sum_{\{a_{i,j}\}} P(\mathbf{G}|\{h_i\}) e^{\mu \sum_{i < j} a_{i,j}} \quad (6)$$

and the average number of links $\langle L \rangle$ is given by

$$\langle L \rangle = \frac{\partial \ln Z(\mu)}{\partial \mu}. \quad (7)$$

By performing the sum over the $\{a_{i,j}\}$ we obtain

$$Z(\mu) = \int \prod_i dh_i e^{-G(\{h_i\})}, \quad (8)$$

where

$$-G(\{h_i\}) = \sum_i \ln[\rho(h_i)] + \sum_{i < j} \ln[p(h_i, h_j) e^{\mu} + 1 - p(h_i, h_j)] \quad (9)$$

For the sake of clarity, we will suppose to deal with a discrete distribution for the h 's, on a finite set \mathcal{D} , and to stress this, we will write ρ_h instead of $\rho(h)$:

$$\sum_{h \in \mathcal{D}} \rho_h = 1. \quad (10)$$

Later on we will extrapolate the continuum limit too. We notice now that $G(\{h_i\})$ depends only on the multiplicities of the h 's:

$$\begin{aligned} -G(\{h_i\}) &= -\tilde{G}(\{N_h\}) \\ &= \sum_h N_h \ln(\rho_h) + \frac{1}{2} \sum_{h, h'} N_h N_{h'} \sigma_{h, h'}, \end{aligned} \quad (11)$$

where $(\delta(x, y))$ stands for Kronecker delta function)

$$N_h = \sum_i \delta(h_i, h), \quad (12)$$

i.e., given any ensemble realization $\{h_i\}$, for any $h \in \mathcal{D}$, N_h provides the number of nodes with hidden variable h , and

$$\sigma_{h, h'} = \ln[p(h, h') e^{\mu} + 1 - p(h, h')] \quad (13)$$

By using Eqs. (9-13), $Z(\mu)$ becomes

$$\begin{aligned} Z(\mu) &= \sum_{\{N_h\}} \mathcal{N}(\{N_h\}) \\ &\times e^{\sum_h N_h \ln(\rho_h) + \frac{1}{2} \sum_{h, h'} N_h N_{h'} \sigma_{h, h'}} \end{aligned} \quad (14)$$

where, $\mathcal{N}(\{N_h\})$ is the number of ways in which we can arrange N nodes such that the multiplicities of the h 's are $\{N_h\}$, and $\sum'_{\{N_h\}}$ stands for sum over all the normalized multiplicities $\sum_h N_h = N$. The number $\mathcal{N}(\{N_h\})$ is given by

$$\mathcal{N}(\{N_h\}) = \frac{N!}{\prod_h N_h!}, \quad (15)$$

therefore, by using the Stirling approximation, Eq. (14) becomes

$$Z(\mu) = \sum_{\{N_h\}}' e^{-F(\{N_h\})}, \quad (16)$$

where now

$$\begin{aligned} -F(\{N_h\}) &= N \ln(N) + \sum_h N_h (\ln(\rho_h) - \ln(N_h)) \\ &+ \frac{1}{2} \sum_{h, h'} N_h N_{h'} \sigma_{h, h'}. \end{aligned} \quad (17)$$

We will evaluate the sum in Eq. (16) by making use of

$$\begin{aligned} \sum_{\{N_h\}}' &\sim \int \prod_h dN_h = \int \prod_h dN_h \delta\left(N - \sum_h N_h\right) \\ &= \frac{1}{2\pi} \int \prod_h dN_h \int dx e^{ix(N - \sum_h N_h)}. \end{aligned} \quad (18)$$

The N_h 's are extensive. Let us rewrite them as

$$N_h = N\nu_h. \quad (19)$$

With these notations we have

$$\begin{aligned} Z(\mu) &= \sum_{\{N_h\}} e^{-F(\{N_h\})} \\ &\sim \frac{N^{|\mathcal{D}|}}{2\pi} \int \prod_h d\nu_h \int dx e^{-Nf(\{\nu_h\};x)}, \end{aligned} \quad (20)$$

where $|\mathcal{D}|$ is the cardinality of the set of the h 's, \mathcal{D} , and

$$\begin{aligned} -f(\{\nu_h\};x) &= \ln(N) \left(1 - \sum_h \nu_h\right) \\ &+ ix \left(1 - \sum_h \nu_h\right) + \sum_h \nu_h (\ln(\rho_h) - \ln(\nu_h)) \\ &+ \frac{1}{2} N \sum_{h,h'} \nu_h \nu_{h'} \sigma_{h,h'}. \end{aligned} \quad (21)$$

Of course, in the continuum limit $|\mathcal{D}| \rightarrow \infty$, so that $Z(\mu)$ is ill defined, but the averages are all well defined. We have

$$\begin{aligned} -\frac{\partial f}{\nu_h} &= -\ln(N) - ix + \ln(\rho_h) - \ln(\nu_h) - 1 \\ &+ N \sum_{h'} \nu_{h'} \sigma_{h,h'}, \end{aligned} \quad (22)$$

and

$$-\frac{\partial f}{\partial x} = i \left(1 - \sum_h \nu_h\right). \quad (23)$$

The saddle point Eqs. $\nabla f = 0$ imply that Eq. (23) fixes the normalization of the ν_h 's: $\sum_h \nu_h = 1$, so that

$$\nu_h = \frac{\rho_h e^{N \sum_{h'} \nu_{h'} \sigma_{h,h'}}}{\sum_{h''} \rho_{h''} e^{N \sum_{h'} \nu_{h'} \sigma_{h'',h'}}}. \quad (24)$$

These Eqs. remain well defined also in the continuum limit $|\mathcal{D}| \rightarrow \infty$, being

$$\nu(h) = \frac{\rho(h) e^{N \int dh' \nu(h') \sigma(h,h')}}{\int dh'' \rho(h'') e^{N \int dh' n(h') \sigma(h,h')}}. \quad (25)$$

Once solved, Eqs. (24) (or Eqs. (25)), give the saddle point solution ν_h^* to be plugged into the expression for $Z(\mu)$, and provide the most likelihood sequences $\{\nu_h^*\}$ that maximizes $\tilde{P}_\mu(\{\nu_h\})$, defined as the probability to pick up at random any graph with given sequence $\{\nu_h\}$ in the presence of a chemical potential μ :

$$\begin{aligned} Z(\mu) &\propto e^{-Nf(\{\nu_h^*\})} \\ &\sim \max_{\{\nu_h\}} \left\{ \mathcal{N}(\{\nu_h\}) \prod_h \rho_h^{N_h} P_\mu(\{N_h\}) \right\} \\ &\propto \max_{\{\nu_h\}} \left\{ \tilde{P}_\mu(\{N_h\}) \right\}, \end{aligned} \quad (26)$$

where (with an abuse of notation, on noticing that $f(\{\nu_h^*\};x)$ does not depend on x , we write $f(\{\nu_h^*\})$ instead of $f(\{\nu_h^*\};x)$)

$$\begin{aligned} -f(\{\nu_h^*\}) &= \sum_h \nu_h^* (\ln(\rho_h) - \ln(\nu_h^*)) \\ &+ \frac{1}{2} N \sum_{h,h'} \nu_h^* \nu_{h'}^* \sigma_{h,h'}. \end{aligned} \quad (27)$$

Numerically, we can solve Eqs. (24) only for discrete h 's. In some cases we can manage also the continuum limit analytically.

C. The role of the chemical potential μ ; complete system of Eqs.

From Eq. (13) we see that

$$\text{sgn}(\sigma_{h,h'}) = \text{sgn}(\mu), \quad (28)$$

hence, when $\mu \rightarrow 0$, $\sigma_{h,h'} \rightarrow 0$ and, as a consequence, Eqs. (24) give $\nu_h^* \rightarrow \rho_h$. As we mentioned in Sec. II, in the case of the scale-free model (1)-(2), we can control the average connectivity $\langle k \rangle = 2\langle L \rangle / N$ via the parameter \bar{k} , by simply choosing $\bar{k} = \langle k \rangle$, and we do not need to operate with a $\mu \neq 0$, so that $\nu_h^* = \rho_h$, consistently with the known fact that the distribution of the expected degree h 's is equal to ρ_h [10-13].

However, if we consider more general cases as, like *e.g.*, RGG, $p(h_i, h_j)$ is a “rigid” or “soft” function of the relative distances between nodes i and j , and we are not able to directly control $\langle k \rangle$ from the parameters of $p(h, h')$. More precisely, we cannot reach the desired $\langle k \rangle$ by only tuning the parameters in $p(h, h')$ (for example by using the sole Eq. (32)). We are then forced to use a finite chemical potential μ . If $\mu \neq 0$, we have $\sigma_{h,h'} \neq 0$ and then $\nu_h \neq \rho_h$; *i.e.*, the system is not in equilibrium with respect to ρ_h . To evaluate μ , from Eqs. (7), (13) and (24), we have

$$\begin{aligned} \langle L \rangle &= -N \frac{\partial f(\{\nu_h^*\})}{\partial \mu}, \quad \text{where} \\ -\frac{\partial f}{\partial \mu} &= \sum_h \frac{-\partial f}{\partial \nu_h} \frac{\partial \nu_h}{\partial \mu} + \frac{1}{2} N \sum_{h,h'} \nu_h \nu_{h'} \frac{\partial \sigma_{h,h'}}{\partial \mu}, \end{aligned}$$

which, when calculated at the saddle point $\nu_h = \nu_h^*$, simplifies as

$$\begin{aligned} -\frac{\partial f(\{\nu_h^*\})}{\partial \mu} &= \frac{1}{2} N \sum_{h,h'} \nu_h^* \nu_{h'}^* \frac{\partial \sigma_{h,h'}}{\partial \mu}, \quad \text{with} \\ \frac{\partial \sigma_{h,h'}}{\partial \mu} &= p_{h,h'} e^{\mu - \sigma_{h,h'}}. \end{aligned}$$

In conclusion, we have to solve the systems of Eqs.

$$\nu_h = \frac{\rho_h e^{N \sum_{h'} \nu_{h'} \sigma_{h,h'}}}{\sum_{h''} \rho_{h''} e^{N \sum_{h'} \nu_{h'} \sigma_{h'',h'}}}, \quad (29a)$$

$$\langle L \rangle = \frac{1}{2} N^2 \sum_{h,h'} \nu_h \nu_{h'} p_{h,h'} e^{\mu - \sigma_{h,h'}}. \quad (29b)$$

or, when \mathcal{D} is continuum

$$\nu(h) = \frac{\rho(h) e^{N \int dh' \nu(h') \sigma(h,h')}}{\int dh'' \rho(h'') e^{N \int dh' \nu(h') \sigma(h'',h')}}}, \quad (30a)$$

$$\langle L \rangle = \frac{1}{2} N^2 \int dh dh' \nu(h) \nu(h') p(h,h') e^{\mu - \sigma(h,h')}. \quad (30b)$$

The sums in (29) and the integrals in (30) are understood to run over the set \mathcal{D} . In the continuum, the system (30) is a functional system. In such a case the discrete system (29) provides a numerical approximation to (30) whose accuracy grows with the number of points N_s used to discretize \mathcal{D} . The system (29) consists of $|\mathcal{D}| + 1$ Eqs. in the $|\mathcal{D}| + 1$ unknowns ($\{\nu_h\}; \mu$), with the $|\mathcal{D}| + 1$ inputs ($\{\rho_h\}; \langle L \rangle$). In the following, with some abuse of notation, sometimes we will write $|\mathcal{D}| = N_s$. It will be clear from the context whether we are referring to a model with discrete value hidden variables, or to the discrete approximation of a model in the continuum.

It is useful to observe that Eq. (30b) (and similarly Eq. (29b)) can also be written as

$$\langle L \rangle = \frac{1}{2} N^2 \int dh dh' \nu(h) \nu(h') q_\mu(h,h'), \quad (31)$$

where we have introduced the probability

$$\begin{aligned} q_\mu(h,h') &= p(h,h') e^{\mu - \sigma(h,h')} \\ &= \frac{p(h,h') e^\mu}{p(h,h') e^\mu + 1 - p(h,h')}. \end{aligned} \quad (32)$$

Eqs. (29)-(32) make clear that we are dealing with an ensemble duality:

$$\text{Ensemble } \{\mathbf{p}, \boldsymbol{\rho}, \mu\} = \text{Ensemble } \{\mathbf{q}_\mu, \boldsymbol{\nu}_\mu, 0\} \quad (33)$$

The duality (33) says that, given N and \mathcal{D} , in order to draw a network with a desired average connectivity $\langle k \rangle = 2\langle L \rangle / N$, we need to look for a proper distribution of nodes $\boldsymbol{\nu}_\mu(h)$ and for a proper distortion of $p(h,h')$ via $q_\mu(h,h')$ to favor ($\mu > 0$) or unfavor ($\mu < 0$) the probability to have a link between h and h' . Since (33) has been obtained by using a saddle point technique, it is exact only in the thermodynamic limit $N \rightarrow \infty$, and the $\text{Ensemble } \{\mathbf{q}_\mu, \boldsymbol{\nu}_\mu, 0\}$ contains only the typical graphs of the $\text{Ensemble } \{\mathbf{p}, \boldsymbol{\rho}, \mu\}$.

IV. A TOY EXAMPLE

Let us consider the classical random graph. Within the hidden-variables formalism one has

$$p(h,h') = \frac{c}{N}, \quad (34)$$

$$\rho(h) = 1, \quad h \in \mathcal{D}, \quad (35)$$

where $\mathcal{D} = [0,1]$ is an interval and $c > 0$ is a constant. Due to the fact that $p(h,h') = \text{constant}$, we have also $\sigma(h,h') = \text{constant}$ and Eqs. (30a)-(30b) simply give

$$\nu_\mu(h) = \rho(h), \quad (36a)$$

$$\langle k \rangle = \frac{ce^\mu}{1 + \frac{c}{N}(e^\mu - 1)} \simeq ce^\mu. \quad (36b)$$

Eqs. (36a)-(36b) tell us that we can reach any desired average connectivity $\langle k \rangle$ without changing $\rho(h)$, and that $\mu \neq 0$ simply amounts to a renormalization of the constant c . Therefore, in this example the duality relation (33) does not add any new information.

V. APPLICATION TO RANDOM GEOMETRIC GRAPHS

A. General definition of RGG

In RGG the h 's represent the positions of the nodes and two nodes are connected or not according to

$$p_{h,h'} = \begin{cases} 1, & \text{if } h \text{ and } h' \text{ satisfy a geometric rule,} \\ 0, & \text{otherwise.} \end{cases} \quad (37)$$

Eqs. (37) plugged into Eqs. (13) give

$$\sigma_{h,h'} = \begin{cases} \mu, & \text{if } h \text{ and } h' \text{ satisfy a geometric rule,} \\ 0, & \text{otherwise.} \end{cases} \quad (38)$$

Eq. (37) represents rigid RGG. More in general, we might consider soft RGG where Eq. (37) is replaced by

$$p_{h,h'} = \begin{cases} a, & \text{if } h \text{ and } h' \text{ satisfy a geometric rule,} \\ 1 - a, & \text{otherwise,} \end{cases} \quad (39)$$

where $0 < a < 1$ can be a constant or any function of the node positions h and h' .

Particularly important are the RGG defined over a continuous domain \mathcal{D} equipped with a distance. In the rigid case ($a = 1$ in Eq. 39), for any sprinkle of nodes, there is one single graph obtained by connecting all pairs of nodes whose distance is at most r , $r > 0$ being a fixed parameter, called threshold. Nodes are sprinkled according to the given PDF $\rho(h)$, and the support of $\rho(h)$ defines the domain \mathcal{D} . A natural choice for $\rho(h)$ is the uniform PDF $\rho(h) = \text{constant}$, and this will be the case in the following numerical examples. However, for practical and theoretical reasons it is convenient to think of $\rho(h)$ as an arbitrary PDF.

We stress that, for rigid RGG we have $q_\mu(h,h') = p(h,h')$ but $\nu_\mu(h) \neq \rho(h)$.

B. Solving the Saddle Point Eqs; Dynamics of RGG

Suppose for simplicity that μ is known. Given any ρ_h , which plays the role of an initial density $\nu_h(t=0) = \rho_h$, Eqs. (29a) leads to a natural dynamics over the ν_h 's. For example, we can define a discrete-time dynamics by simple iteration of the ν_h 's:

$$\begin{aligned}\nu_h(t+1) &= \frac{\rho_h e^{N \sum_{h'} \nu_{h'}(t) \sigma_{h,h'}}}{\sum_{h''} \rho_{h''} e^{N \sum_{h'} \nu_{h'}(t) \sigma_{h'',h'}}}, \\ \nu_h(t=0) &= \rho_h.\end{aligned}\quad (40)$$

Under the hypothesis that there exists a stable fixed point, the above iteration offers a way to solve Eqs. (29a) iteratively. There are many ways to define the dynamics, but all make use of Eqs. (29a) and therefore have the same stationary solutions. In general, if $\mu \neq 0$, Eqs. (29a) can generate a “diffusion” with asymptotic solution toward $\nu_h(t=\infty) = \nu_h$. This picture is very appealing for RGG. However, we have to keep in mind that we need to work with a large value of $|\mathcal{D}| = N_s$ in order to recover a diffusion-like trend. In fact, if we choose for ρ_h a point-like distribution:

$$\rho_h = \delta_{h,h_0}, \quad (41)$$

it is immediate to check that Eqs. (29a) are solved with

$$\nu_h = \delta_{h,h_0}. \quad (42)$$

More in general, ν_h is always of the form

$$\nu_h = \chi_{\mathcal{D}}(h) f_h, \quad (43)$$

where $\chi_{\mathcal{D}}(h)$ is the characteristic function of the set \mathcal{D} and f_h a suitable vector. It is then clear that for small values of $|\mathcal{D}|$, we cannot have a diffusion like trend. We expect to see a diffusion when $|\mathcal{D}| \sim N^\alpha$, with $\alpha > 0$. RGGs, however, are usually defined over a continuous domain \mathcal{D} and for an exact solution we should solve the functional system (30). As anticipated, a numerical solution relies on a discretization of \mathcal{D} by using a large enough number of points N_s . Given N , the numerical solution will approach the exact solution in the limit $N_s/N \rightarrow \infty$. However, for all practical aims a finite ratio provides excellent approximations. Note that solving (29) by using the iteration (40) produces only the solutions that are minima (local or global) of the free energy density (28).

Finally, we point out that, although there might exist smarter methods to solve (29), traditional population dynamics here cannot be used. In fact, in population dynamics the iteration does not run for any h and the sums in (40) are sampled for randomly selected sites and their respective neighbors (here two sites are neighbors if $\sigma_{h,h'} \neq 0$). Population dynamics returns the exact iteration when $N_s \rightarrow \infty$. However, it should be clear that such a sampling relies on the statistical knowledge of the first neighbors of each site that must be finite. In

our case we do not have access to this knowledge when $N_s \rightarrow \infty$. In fact, if \mathcal{D} is continuum, given h , the number of site-neighbors of h diverges for $N_s \rightarrow \infty$. In conclusion, we are forced to rely to the full exact iteration (40) with finite N_s .

C. RGG on the circle

We can deepen our understanding with the simplest RGG: the circle. In this case $\mathcal{D} = [0, 2\pi)$. Given the number of nodes N , we choose a discretization of \mathcal{D} with N_s equidistant sites; in other words the N nodes can occupy any of the N_s sites whose positions are given by their angles θ_h , $h = 1, \dots, N_s$. The N nodes are randomly sprinkled through a given distribution ρ_h . Then we connect two points according to the following probability

$$p(\theta_h, \theta_{h'}) = \begin{cases} a, & \|\theta_h - \theta_{h'}\| < r, \\ 1-a, & \|\theta_h - \theta_{h'}\| \geq r, \end{cases} \quad (44)$$

where a and r are constant, and $\|\theta - \theta'\| = \pi - |\pi - |\theta - \theta'||$ is the distance on the circle. We set

$$r = \frac{4\pi}{N_s}. \quad (45)$$

Eqs. (44) plugged into Eqs. (13) give

$$\sigma(\theta_h, \theta_{h'}) = \begin{cases} \ln[ae^\mu + 1 - a], & \|\theta_h - \theta_{h'}\| < r, \\ \ln[(1-a)e^\mu + a], & \|\theta_h - \theta_{h'}\| \geq r. \end{cases} \quad (46)$$

If for ρ_h we choose the uniform distribution

$$\begin{aligned}\rho_h &= \rho(\theta_h) = \frac{1}{N_s}, \\ \theta_h &\in \left\{0, \frac{2\pi}{N_s}, \frac{4\pi}{N_s}, \dots, \frac{(N_s-1)2\pi}{N_s}\right\},\end{aligned} \quad (47)$$

it is immediate to check that $\nu_h = \rho(\theta_h)$, independently from μ and a (this holds for any homogeneous RGG: if ρ_h is the uniform distribution, $\nu_h = \rho_h$). However, for $\rho(\theta_h)$ we can choose a distribution which differs slightly from the uniform one:

$$\rho_h = A(1 + \epsilon_h) \quad (48)$$

where ϵ_h are uniformly distributed random variables with mean 0 and finite variance, and A is a normalization constant. In general, if we use Eq. (40) with the initial distribution (48), the system may evolve toward a distribution which is different from the uniform one. In principle, any initial distribution (provided not equal to the uniform one) can be used to solve the system (29). In fact, we find that via (40) they all tend to the same solution for (29), provided they have the same support in \mathcal{D} .

We expect that, the larger is $\langle k \rangle = \langle 2L \rangle / N$, the larger is μ . However, it is a priori difficult to know whether, for any $\langle k \rangle$, there corresponds a solution for (29). In fact,

in general this is not the case, furthermore, we find out that $\langle k \rangle$, as a function of μ , undergoes a $O(N)$ jump in correspondence of a critical μ_c , as depicted in Figs. 2-4 where we analyze several values of a .

D. The general scenario: a geometry-induced condensation of nodes and links

The phase transition depicted in Figs. 1-4 is perfectly compatible with the geometric interpretation and we can summarize the general scenario as follows. Given N nodes and a rigid ($a = 1$) RGG defined over a continuous domain \mathcal{D} , we are free to arrange the N nodes in any way, with two opposite limit regimes: If nodes are uniformly distributed over \mathcal{D} , $\langle k \rangle = O(1)$, while if nodes are localized in a small subset of \mathcal{D} , $\langle k \rangle = O(N)$. However, our analysis shows that, when $N \rightarrow \infty$, the probability to find a configuration out of these two regimes tends to 0, and the two regimes become two phases separated by a first order phase transition: a uniformly diluted/regular phase, and a condensate of nodes and links in the other phase. Such a transition is triggered by the geometry whose strength can be tuned by the value of the parameter $0 < a \leq 1$. For $a > 1/2$ the two regimes are separated by the geometry: Close nodes have an high number of links. For $a < 1/2$ the two regimes are separated by an anti-geometry: Close nodes have a low number of links. Finally, for $a = 1/2$ there is no geometry and no phase transition. See Fig. 5 for a qualitative phase diagram description. We stress that this phase transition scenario has nothing to share with the percolation phenomena (in Appendix B we show how to deal with the percolation problem within our framework). In fact, as we shall prove later by a combinatorial argument, the scenario does not depend on the details of the model, nor on the dimension d , or on the particular value r chosen, which can be changed to scale in the very sparse or dense regimes (as we have numerically checked by replacing r in Eq. (45) with $r \propto 1/N^2$ and $r \propto \ln(N)/N$, respectively) without affecting the qualitative features of the phase transition.

E. Combinatorial argument

From a graph construction viewpoint, one might wonder why there is this jump in $\langle k \rangle$. In other words, if we are free to lay down nodes as we like, why are only certain values of $\langle k \rangle$ visible when $N \rightarrow \infty$? Given a RGG with domain \mathcal{D} , let us consider for simplicity the rigid case $a = 1$. In this case, given the connectivity $\langle k \rangle$ (or, equivalently, given μ), the node-positions are all equiprobable. It is also convenient to proceed with a discretized version of \mathcal{D} , with $|\mathcal{D}| = N_s = \alpha N$, where $\alpha \gg 1$ is an arbitrarily large constant. When we look at all the possible realizations of the graphs, we start with the graph with minimal connectivity $\langle k \rangle = O(1)$, where the nodes are uniformly distributed and this configuration is

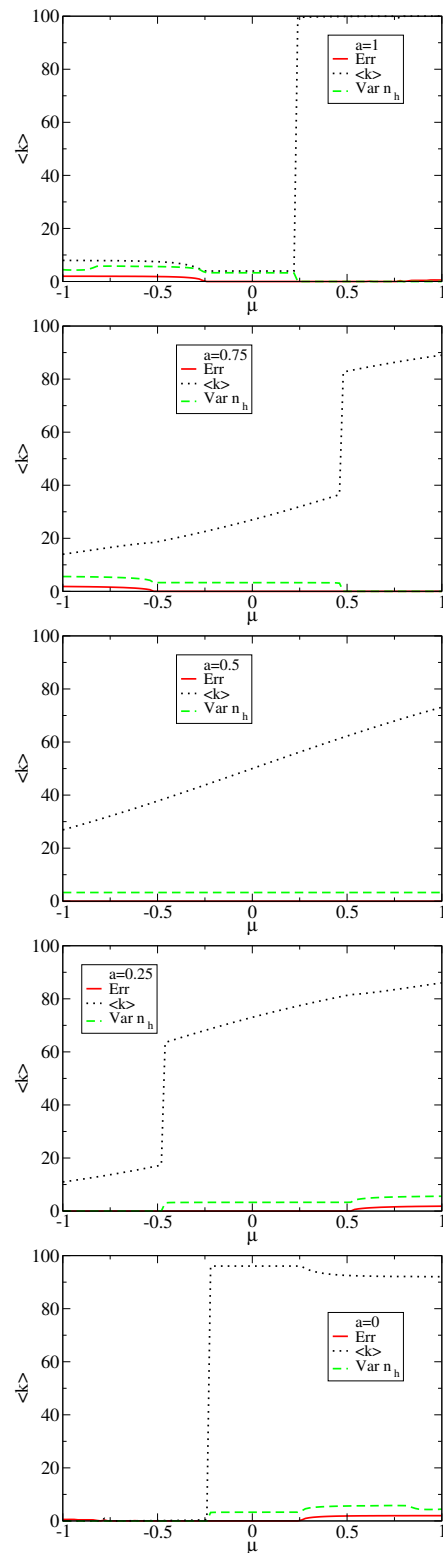


FIG. 2: (Color online) Analysis of the RGG (44)-(45) for several values of a (here ν_h stands for ν_θ). We plot $\langle k \rangle$ vs the chemical potential μ . $\text{Err} = \sum_h |\nu_h(t+1) - \nu_h(t)|$ provides an evaluation of the existence of solutions: For $\text{Err} > 0$ there are no solutions; $\text{Var } \nu_h$ is the variance of $\nu_h(t)$. Data obtained by solving the discrete system (29) via the iteration (40) with $t = 200$ iterations. Here the number of nodes is $N = 100$, whereas the discretization parameter is set to $N_s = 1000$. The RGG is exactly reproduced only in the limit $N_s \rightarrow \infty$, however, for our numerical aims $N_s = 1000$ is large enough and, on this scale, larger values of N_s do not produce serious differences.

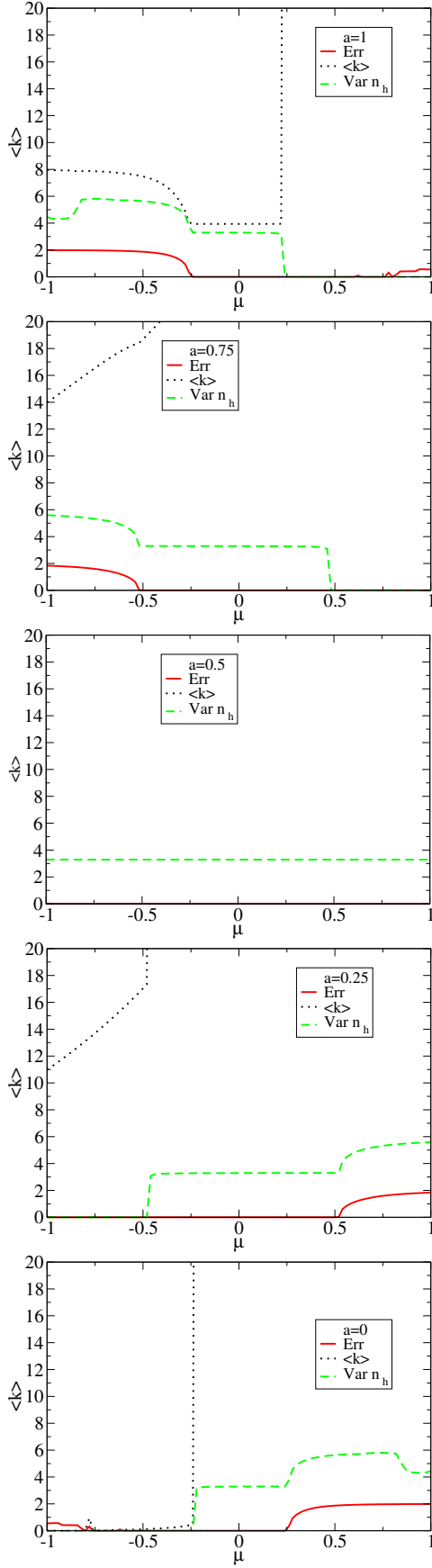


FIG. 3: (Color online) Analysis of the RGG (44)-(45) for several values of a (here n_h stands for ν_θ). Enlargements of the plots of Figs. 2.

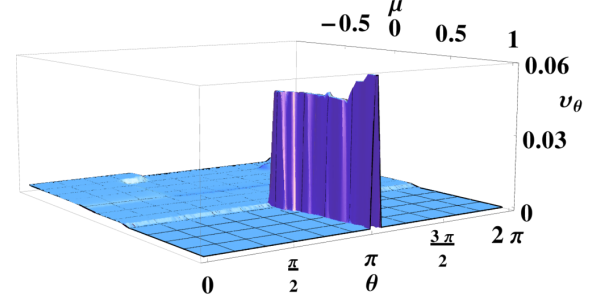


FIG. 4: (Color online) Distribution ν_θ for the RGG (44)-(45) with $a = 1$, as a function of the angle θ for each value of the chemical potential μ . Plot obtained with $N = 100$, $N_s = 1000$ and $t = t = 200$, as for Fig. 2. Note that, in the localized regime, there is an high degeneracy of solutions, since the N points can get localized in any small region of the interval $[0, 2\pi)$. Up to this degeneracy, different initial conditions (except the trivial one (41)) lead to the same solution. Finite size effects can be seen in Figs. 1.

roughly unique (0 degeneracy). Next, we consider the opposite case where the graph has maximal connectivity $\langle k \rangle = N$, where the nodes are densely localized over a small region (average distance smaller than r), and this configuration has degeneracy which goes like N . Let us now consider the configurations in which there are two separated highly dense regions having $N/2$ nodes each. In this latter case we have $\langle k \rangle = N/2$, and the degeneracy of such possible configurations goes like N^2 . According to this counting, we should then expect to see this latter case, where $\langle k \rangle = N/2$, as highly more probable than the former case, where $\langle k \rangle = N$, which is the opposite of what we have seen in the previous paragraph (see top panel of fig. 2). The apparent paradox is due to the fact that such a naive counting does not take into account the perturbations from these ideal cases. Suppose we perturb slightly the case with $\langle k \rangle = N$ in the following way: We lay down all the nodes again in a single small region, yet this region has a slightly larger extension such that the total number of links is $L_1 = x \cdot N_1$, where $N_1 = \binom{N}{2}$, and where $0 < x < 1$, with $x \sim 1$. Due to the fact that N_s can be set as large as we like (\mathcal{D} is continuous), this is always possible, for any x . When $x \rightarrow 1$ we recover the ideal case having only one possible type of realization: the fully connected graph. However, when $x < 1$, the number of possible graphs \mathcal{N}_1 that we can construct in this way goes like

$$\mathcal{N}_1(x) = \alpha N \binom{N_1}{L_1} \simeq \alpha N e^{N_1 s(x)},$$

$$\text{with } \langle k \rangle = \frac{2L_1}{N} = Nx \quad (49)$$

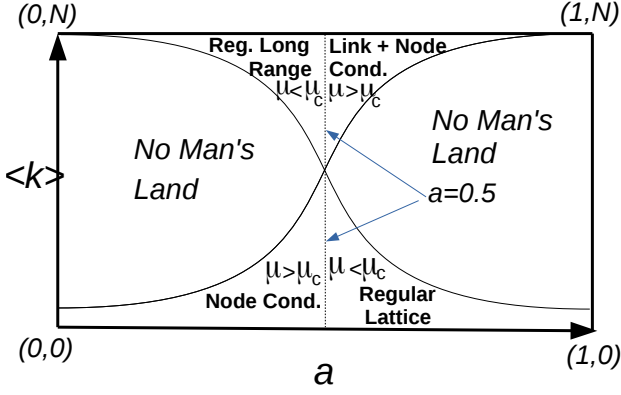


FIG. 5: (Color online) Pictorial phase diagram of a Euclidean RGG in the plane $(a, \langle k \rangle)$. Given a threshold r , two nodes are connected with probability a if their distance is at most r . The regions labeled “No Man’s Lands” are regions populated by only atypical (*i.e.*, exponentially rare, or hard solutions) graphs. There are four stable phases (or easy solutions): a condensate and a regular region, for both $a > 1/2$ and $a < 1/2$. Case $a > 1/2$ (close nodes are connected with probability $a > 1/2$): the condensate zone is the region where typical graphs are characterized by a condensation of nodes and links, with the resulting graphs nearly fully connected, whereas the regular zone is the region where typical graphs have nodes nearly uniformly distributed and with a minimal connectivity, as in a regular lattice. Condensate and regular regions are characterized by $\mu > \mu_c$ and $\mu < \mu_c$, respectively. For $a < 1/2$ (distant nodes are connected with probability $1 - a > 1/2$) the picture is reversed, with condensed nodes having minimal connectivity, and uniformly distributed nodes having maximal connectivity (due to long range connections), for $\mu > \mu_c$ and $\mu < \mu_c$, respectively. At $a = 0.5$ there is no geometry and no phase transition. Notice that μ_c depends on a . We point out also that, in the exact phase diagram, the minimal and maximal average connectivities of the cases $a > 0.5$ are slightly smaller than the minimal and maximal average connectivities of the cases $a < 0.5$. The reason for this difference is due to a simple geometrical argument whose details are left to the reader.

where the approximation holds for N large, and $s(x) \geq 0$ is a “one-particle entropic factor”

$$s(x) = -\ln(x)x - (1-x)\ln(1-x). \quad (50)$$

Similarly, if we perturb slightly the ideal case with $\langle k \rangle = N/2$, in at least one of the two highly dense and separated regions, by introducing $N_2 = \binom{N/2}{2}$, and $L_2 = x \cdot N_2$, we see that the total number of possible graphs \mathcal{N}_2 that we can construct in this way goes like

$$\mathcal{N}_2(x) = \binom{\alpha N}{2} \binom{N_2}{L_2} \simeq \binom{\alpha N}{2} e^{2N_2 s(x)},$$

with $\langle k \rangle = \frac{2L_2}{N} = \frac{N}{2}x$, (51)

where $s(x)$ is again given by Eq. (50). If we put at ratio Eqs. (49) and (51) we get

$$\frac{\mathcal{N}_2(x)}{\mathcal{N}_1(x)} \simeq \frac{\alpha N}{2} e^{(2N_2 - N_1)s(x)} = \frac{\alpha N}{2} e^{\frac{-N^2}{2}s(x)}. \quad (52)$$

We can iterate the above argument for other values of $\langle k \rangle$. In particular, we can consider the configurations having 2^q small equally populated regions, each having an average connectivity $\langle k \rangle = N/2^q$, where q is an integer. In this case the leading term of the ratio between $\mathcal{N}_1(x)$ and $\mathcal{N}_{2^q}(x)$ goes like

$$\frac{\mathcal{N}_{2^q}(x)}{\mathcal{N}_1(x)} \sim \left(\frac{\alpha N}{2^{q-1}} \right) e^{-\frac{N^2}{2^q}s(x)}, \quad (53)$$

It is then clear that, as far as q is finite, the configurations with $\langle k \rangle \simeq N$ highly dominate the others with $\langle k \rangle \simeq N/2^q$. We might have a different situation only when $q = O(\ln(N))$, where $\langle k \rangle = O(1)$, consistently with what we have shown numerically in the previous paragraphs. Similar arguments hold for other intermediate values of $\langle k \rangle$.

Remarkably, this combinatorial argument does not depend on the details of the RGG, or on the dimension d , or the nature of the underlying geometry, which, in principle, can also be non Euclidean. As soon as we are dealing with a continuous set \mathcal{D} equipped with some distance, our combinatorial argument can be equally applied.

F. An easy-hard-easy transition

We can better formalize the findings of the previous paragraphs as follows. Let us consider a rigid ($a = 1$) d -dimensional Euclidean RGG of total volume 1 and threshold $r \propto N^{-1/d}$ (see Sec. V.A). For finite N and N_s , let $\mathcal{N}(\langle k \rangle)$ be the number of graphs having average connectivity $\langle k \rangle$ each, and let $\mathcal{N} = \sum_{\langle k \rangle} \mathcal{N}(\langle k \rangle)$ be the total number of graphs (for N_s finite the sum runs over a finite set of possible values of $\langle k \rangle$). If we introduce the PDF

$$f(\langle k \rangle) = \lim_{N_s \rightarrow \infty} \frac{\mathcal{N}(\langle k \rangle)}{\mathcal{N}}, \quad (54)$$

our approach shows that, for N large

$$f(\langle k \rangle) \rightarrow b\delta(\langle k \rangle - \Omega_d r^d (N-1)) + (1-b)\delta(\langle k \rangle - (N-1)), \quad (55)$$

where $\langle k \rangle = \Omega_d r^d (N-1) = O(1)$ corresponds to the connectivity of a regular d -dimensional lattice, $\langle k \rangle = N-1$ with the connectivity of a fully connected graph, $\delta(x)$ stands for Dirac delta function, and $0 < b < 1$ is a constant. Eq. (55) is consequence of the first order phase transition in $\langle k \rangle$ piloted by μ . On the other hand, the combinatorial argument of the previous paragraph leads us to interpret the phase transition as an easy-hard-easy

transition. In fact, when we want to draw graphs with a very low connectivity, $\langle k \rangle = O(1)$, or a very high connectivity, $\langle k \rangle = O(N)$, we are able to figure out how to locate the N nodes in order to have such connectivities. These two opposite cases are indeed relatively easy to build, and their extreme ideal limits corresponding to $\langle k \rangle = \Omega_d r^d (N - 1)$, and to $\langle k \rangle = N - 1$, are even trivial. Note that, as the combinatorial argument makes clear, we are still able to build graphs with any desired connectivity $\langle k \rangle$. In order to do so, it is in fact enough to split the N nodes into $m = N/(\langle k \rangle + 1)$ groups each containing $\langle k \rangle + 1$ nodes which, inside the group, are sufficiently close to each other so that each forms a fully connected graph of $\langle k \rangle + 1$ nodes. If the ratio $m = N/(\langle k \rangle + 1)$ is not an integer, it is still possible to consider slight variations of this construction in order to reach the desired average connectivity (for example by introducing some asymmetry among the groups). However, such solutions are very specific. An algorithm that aims at finding the distribution of nodes producing the target $\langle k \rangle$, and that is based on a uniform sampling over the node positions, would hardly converge (in a time that scales polynomially with N) if $\langle k \rangle \neq \Omega_d r^d (N - 1)$ or $\langle k \rangle \neq N - 1$. In fact, our approach shows that we are in the presence of an easy-hard-easy transition: the first order phase transition marks the boundary between two easy phases between which, in correspondence of the critical point μ_c , there are the hard solutions. Furthermore, the above specific solution with m groups has the bad aspect to be in general a disconnected graph (notice that $m < N$). In fact, in most of practical problems, like in ad hoc mobile-networks, one is interested to find solutions which are also connected graphs. Our guess is that such hard connected solutions correspond to specific heterogeneous graphs. Chosen any point of \mathcal{D} as a reference center, in these solutions, the highest connected nodes are located near the center of the domain \mathcal{D} , whereas the lowest connected nodes are located near the boundaries of \mathcal{D} . The idea is that, radially, nodes are distributed according to a density that decays exponentially with the radial distance with some exponent α . The larger is α , the smaller is the region occupied by the nodes and - as a consequence - the larger is $\langle k \rangle$. This construction is similar to the construction of hyperbolic RGG where (embedded in a hyperbolic space) nodes are instead uniformly distributed (with respect to the hyperbolic metric) [25]. The network so constructed is connected and can give rise to a wide spectra of cases according to γ - the exponent characterizing the degree distribution $p(k)$ - which in turn depends on α , but the network could also be non power law. Let us consider a $d = 1$ dimensional RGG with $\mathcal{D} = [0, 1]$ and threshold r . For a normalized exponential PDF we have

$$\nu(x) = \frac{\alpha}{1 - e^{-\alpha}} e^{-\alpha x}. \quad (56)$$

It is easy to see that, given N nodes, the expected degree $\bar{k}(x)$ of a node located at x is

$$\bar{k}(x) = \frac{N e^{-\alpha x}}{1 - e^{-\alpha}} 2 \sinh(\alpha r), \quad (57)$$

therefore, for $\langle k \rangle$ we have

$$\langle k \rangle = \frac{1 - e^{-2\alpha}}{(1 - e^{-\alpha})^2} N \sinh(\alpha r). \quad (58)$$

Similar expressions hold in any dimension d for a PDF $\nu(\mathbf{x})$ that decays exponentially with the radial distance: $\nu(\mathbf{x}) = b \exp(-\alpha|\mathbf{x}|)/\Omega_d$, with b normalization constant. In particular, it is possible to show that $C_1 \exp(-\alpha|\mathbf{x}|) \leq \bar{k}(\mathbf{x}) \leq C_2 \exp(-\alpha|\mathbf{x}|)$, where C_1 and C_2 are two constants. Eq. (58) shows that, whatever r is, by properly choosing α we can achieve any desired $\langle k \rangle$. In particular, if we consider the standard choice $r \propto 1/N$, for $\alpha = O(1)$ we have $\langle k \rangle = O(1)$, while for $\alpha = O(N)$ we have $\langle k \rangle = O(N)$, *i.e.*, the regimes corresponding to the two easy phases. Yet, the number of possible graphs that one can actually build by using this construction strongly depends on α , or on γ , if the resulting degree distribution is scale free. In fact, in [26], Del Genio *et al.* have proved that scale free graphs with $0 < \gamma < 2$ are either very rare or do not exist, while they exist for $\gamma > 2$ and $\gamma < 0$. Graphs with $\gamma > 2$ are networks with $\langle k \rangle = O(1)$, and are in correspondence to one of the two easy phases of our satisfaction problem, whereas graphs with $\gamma < 0$ correspond to graph realizations nearly fully connected, and are in the other easy phase where $\langle k \rangle = O(N - 1)$. Graphs with $0 < \gamma < 2$, if any, are instead networks that would be able to give rise to $\langle k \rangle = O(N^\beta)$ with $0 < \beta < 1$, *i.e.*, graphs in the set of the hard solutions of our satisfaction problem (if not empty), but the result of [26] forbids their existence. On the other hand, Eqs. (56)-(58) are exact and define a way to build graphs for any α and therefore for any desired $\langle k \rangle$: even if the number of such graphs strongly depends on α , it is never zero. It is possible to show that this conclusion does not contradict the finding of [26]. In fact, if we indicate with $p(k|\mathbf{x})$ the conditional probability that a node located at \mathbf{x} has degree k , from $p(k) = \int d\mathbf{x} \nu(\mathbf{x}) p(k|\mathbf{x})$ and $p(k|\mathbf{x}) = (\bar{k}(\mathbf{x}))^k / k! \exp(-\bar{k}(\mathbf{x}))$ [11], by using the asymptotic behavior of the incomplete gamma function, for $d = 1$, $r \propto 1/N$ and N large, one has

$$p(k) \sim \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}, \quad \text{for } \alpha = O(1), \quad (59)$$

$$p(k) \sim \text{Constant}, \quad \text{for } \alpha = O(N^\beta), \quad 1 \geq \beta > 0. \quad (60)$$

Similar expressions can be found as upper bounds for $d > 1$. Eqs. (59) and (60) show that, when we are in the sparse regime with $\langle k \rangle = O(1)$, the hard solutions correspond approximately to a Poissonian $p(k)$, while, when we are in a dense regime, with $\langle k \rangle = O(N^\beta)$, the hard solutions correspond approximately to a uniform $p(k)$,

where all nodes tend to form nearly fully connected structures. The two cases correspond, formally, to $\gamma \rightarrow +\infty$ and to $\gamma \rightarrow 0$, respectively, compatibly with [26].

VI. CONCLUSIONS

By making use of generating function and saddle point techniques, we have derived the equations for the typical node distributions $\nu(h)$ of a generic hidden-variables model. We have then applied these equations to RGG to face a non trivial satisfaction problem: Given N nodes, a domain \mathcal{D} , and a desired average connectivity $\langle k \rangle$, find - if any - the distribution of nodes having support in \mathcal{D} and average connectivity $\langle k \rangle$. However, the numerical solutions of these equations for $d = 1$ Euclidean RGG shows that the typical $\nu(h)$, in the thermodynamic limit, can only be either uniformly distributed or highly condensed, the two regimes being separated by a first order phase transition characterized by a $O(N)$ jump of $\langle k \rangle$. Other intermediate values of $\langle k \rangle$ correspond in fact to very rare graph realizations. We have then provided a combinatorial argument to fully explain the mechanism inducing this phase transition in general RGG and recognize it as an easy-hard-easy transition triggered by the geometry, and that the hard and connected solutions correspond to strongly heterogeneous constructions embedded in the geometrical space, but these are not necessarily scale free. Our result concludes that, in general, ad hoc optimized networks embedded in geometrical spaces can hardly be designed, unless to rely on very specific constructions.

In our approach, a crucial mathematical tool has been the use of a chemical potential μ in order to tune the desired average connectivity $\langle k \rangle$. Similarly, one could include in the approach other free parameters in order to control, for example, the average number of triangles, or other interesting graph metrics. Notice that, once the solution $\nu(h)$ has been found, one has access not only to the averages of a graph metric, but also to its higher moments. It would be interesting, for certain practical problems described via hidden variable models, to investigate how to better exploit this approach, for example, by requiring that some metrics have also minimal fluctuations, or by requiring that certain correlations are reproduced, similarly to the analysis performed in [24], where clustered scale free models are tuned in such a way to reproduce metrics observed in real world networks.

A different urgent question concerns what this phase transition scenario implies for non Euclidean RGG. We have already mentioned the parallelism with the hyperbolic case [25]. Although our combinatorial argument suggests that this phase transition is expected to be present in any RGG defined over a continuous domain, the issue requires further thoughtful studies.

Acknowledgments

Work supported by CNPq Grant PDS 150934/2013-0. We thank D. Krioukov for early discussions.

Appendix A: Absence of typical detailed configurations in hidden-variables models

In this appendix we show the lack of typical detailed configurations in the ensemble in which the hidden-variables are not fixed. Let be given a hidden variable model via the PDF $\rho(h)$ and the link probability $p(h_i, h_j)$. The conditional probability to realize a graph \mathbf{G} with frozen hidden variables $\{h_i\}$ is

$$P(\mathbf{G}|\{h_i\}) = \prod_{i<j} [p(h_i, h_j)]^{a_{i,j}} [1 - p(h_i, h_j)]^{1-a_{i,j}},$$

so that the PDF to have \mathbf{G} together with the hidden-variables values $\{h_i\}$ is

$$P(\mathbf{G} \cap \{h_i\}) = \prod_i \rho(h_i) \prod_{i<j} [p(h_i, h_j)]^{a_{i,j}} \times [1 - p(h_i, h_j)]^{1-a_{i,j}},$$

while the (unconditioned) probability to have \mathbf{G} is

$$P(\mathbf{G}) = \int \prod_i \rho(h_i) \prod_{i<j} [p(h_i, h_j)]^{a_{i,j}} \times [1 - p(h_i, h_j)]^{1-a_{i,j}}.$$

The generating function of these probabilities can be obtained from

$$\begin{aligned} Z(\{x_{i,j}\}; \{y_i\}) &= \int \prod_i dh_i \rho(h_i) e^{y_i h_i} \sum_{\{a_{i,j}\}} P(\mathbf{G}|\{h_i\}) e^{x_{i,j} a_{i,j}} \quad (\text{A1}) \end{aligned}$$

where $\{x_{i,j}\}$ and $\{y_i\}$ are link- and node-auxiliary fields, respectively. Actually, we do not need to use the y 's, since they do not have (at least here) an interesting graph meaning. However, just for completeness, for the moment being we keep the y 's general, while we will set them to 0 later on. Let us rewrite Z as (compact notation: $p_{i,j} = p(h_i, h_j)$)

$$\begin{aligned} Z(\{x_{i,j}\}; \{y_i\}) &= \int \prod_i dh_i \prod_i \rho(h_i) e^{y_i h_i} \\ &\prod_{i<j} \sum_{a_{i,j}=0,1} e^{a_{i,j}(\ln(p_{i,j}) + x_{i,j}) + (1-a_{i,j}) \ln(1-p_{i,j})}. \quad (\text{A2}) \end{aligned}$$

By summing over the $a_{i,j}$, we get

$$Z(\{x_{i,j}\}; \{y_i\}) = \int \prod_i dh_i e^{N f(\{x_{i,j}\}; \{y_i\}; \{h_i\})} \quad (\text{A3})$$

where

$$Nf(\{x_{i,j}\}; \{y_i\}; \{h_i\}) = \sum_i (y_i h_i + \ln(\rho(h_i))) + \sum_{i < j} \ln(1 - p_{i,j} + p_{i,j} e^{x_{i,j}}). \quad (\text{A4})$$

For N large we can try to use a “saddle-point” technique by solving for the h ’s the system of the N coupled Eqs.

$$\frac{\partial f}{\partial h_i} = 0, \quad i = 1, \dots, N. \quad (\text{A5})$$

From Eq. (A4) we have

$$\frac{N\partial f}{\partial h_i} = y_i + \frac{1}{\rho(h_i)} \frac{\partial \rho(h_i)}{\partial h_i} + \sum_{j:j \neq i} \frac{(e^{x_{i,j}} - 1)}{p_{i,j}(e^{x_{i,j}} - 1) + 1} \frac{\partial p_{i,j}}{\partial h_i}. \quad (\text{A6})$$

Let us specialize now for the following family of hidden-variables models:

$$\rho(h) = ah^{-\gamma}, \quad (\text{A7})$$

$$p(h_i, h_j) = h_i h_j w(h_i h_j), \quad (\text{A8})$$

where a is a normalization constant, and $w(x)$ is a positive dumping factor such that $x w(x) \leq 1$ for any $x \geq 0$, which in particular implies $w(x) \rightarrow 0$, $w'(x) \rightarrow 0$, and $w(x)/w'(x) \rightarrow 0$ for $x \rightarrow \infty$. For the model (A7-A8), the derivatives take simple forms:

$$\frac{\partial p_{i,j}}{\partial h_i} = \frac{p_{i,j}}{h_i} + \frac{p_{i,j}^2}{h_i} \frac{w'_{i,j}}{w_{i,j}^2}. \quad (\text{A9})$$

In particular, for the standard case $w(x) = 1/(1+x)$, we have

$$\frac{\partial p_{i,j}}{\partial h_i} = \frac{p_{i,j}}{h_i} - \frac{p_{i,j}^2}{h_i}. \quad (\text{A10})$$

For N large and $\gamma > 2$, we can neglect terms in $p_{i,j}^2$ w.r.t. to those in $p_{i,j}$ (this approximation is however not essential here). In conclusion, for the model (A7-A8), Eqs. (A6) become

$$\frac{N\partial f}{\partial h_i} = y_i - \frac{\gamma}{h_i} + \frac{1}{h_i} \sum_{j:j \neq i} \frac{p_{i,j}(e^{x_{i,j}} - 1)}{p_{i,j}(e^{x_{i,j}} - 1) + 1}, \quad (\text{A11})$$

which leads to the following system of saddle-point Eqs. for the h ’s

$$y_i h_i = \gamma - \sum_{j:j \neq i} \frac{p_{i,j}(e^{x_{i,j}} - 1)}{p_{i,j}(e^{x_{i,j}} - 1) + 1}, \quad i = 1, \dots, N. \quad (\text{A12})$$

For a given set of values of the auxiliary fields $\{x_{i,j}\}; \{y_i\}$, Eqs. (A12) can have one or more solutions. We will indicate a solution of the saddle-point Eqs. (A12) with

the superscript $*$: $\{h_i^*\}$. Note that $h_i^* = h_i^*(\{x_{i,j}\}; \{y_i\})$. If we set $\{y_i = 0\}$, Eqs. (A12) reduce to

$$\gamma = \sum_{j:j \neq i} \frac{p_{i,j}(e^{x_{i,j}} - 1)}{p_{i,j}(e^{x_{i,j}} - 1) + 1}, \quad i = 1, \dots, N. \quad (\text{A13})$$

However, we immediately see that Eqs. (A13) have no solution for $\{x_{i,j} = 0\}$, which is the value of the auxiliary fields where we have to set up our calculations at the end to get Z , and its derivatives (in order to get also the averages). In other words, in the ensemble where the h ’s are random variables there are no typical graphs.

We can alternatively try to make first the integral over the h ’s and only later to sum over the $\{a_{i,j}\}$. We get

$$Z(\{x_{i,j}\}) = \sum_{\{a_{i,j}\}} \int \prod_i dh_i e^{Ng(\{x_{i,j}\}; \{a_{i,j}\}; \{h_i\})}, \quad (\text{A14})$$

where

$$Ng(\{x_{i,j}\}; \{a_{i,j}\}; \{h_i\}) = \sum_i \ln(\rho(h_i)) + \sum_{i < j} [a_{i,j}(\ln(p_{i,j}) + x_{i,j}) + (1 - a_{i,j}) \ln(1 - p_{i,j})]. \quad (\text{A15})$$

We have

$$\frac{N\partial g}{\partial h_i} = -\frac{\gamma}{h_i} + \frac{1}{h_i} \sum_{j:j \neq i} \frac{p_{i,j}(e^{x_{i,j}} - 1)}{p_{i,j}(e^{x_{i,j}} - 1) + 1}, \quad (\text{A16})$$

which leads to the following system of saddle-point Eqs. for the h ’s

$$\gamma = \sum_{j:j \neq i} \frac{a_{i,j} - p_{i,j}}{1 - p_{i,j}}, \quad i = 1, \dots, N. \quad (\text{A17})$$

It is not possible to satisfy such saddle-point Eqs..

Appendix B: Percolation in RGG

Percolation in RGG has been studied for Euclidean RGG’s and also rigorously in $d = 1$ and $d = 2$ dimensions [14, 16]. It is not the aim of this paper to analyze in detail the percolation in RGG, however, it should be clear that the solution $\nu(h)$ of the saddle point Eqs. (29a) contains crucial information about the percolation problem. In particular, we can analyze the percolation as follows. Let us consider a d -dimensional rigid ($a = 1$) Euclidean RGG with domain $\mathcal{D} = [0, 1)^d$. Here h represents a position vector in $[0, 1)^d$. Given $r > 0$, two nodes are connected if and only if their euclidean distance is at most r . Given the number of nodes N , the initial density $\rho(h)$ (which in turn defines \mathcal{D}), and the solution $\nu(h)$ of Eqs. (29a), let us define the following set in the continuum

$$\mathcal{D}_1 = \{h \in \mathcal{D} : N \int_{\mathcal{B}(h,r)} \nu(h') dh' \geq 1\}, \quad (\text{B1})$$

where $\mathcal{B}(h, r)$ stands for the d -ball of radius r centered at h . The set \mathcal{D}_1 provides the positions h 's where, in average, there is at least one node within the balls $\mathcal{B}(h, r)$'s. It is then clear that, in the limit in which we can neglect fluctuations of the node positions, the RGG will be percolating if \mathcal{D}_1 is a connected set in $\mathcal{D} = [0, 1]^d$ (in the topological sense). Let us consider the standard case where nodes are uniformly sprinkled over $[0, 1]^d$, *i.e.*, $\rho(h) = 1$. In this case we have also $\nu(h) = 1$, from which we get that, depending on the value of r , we have either $\mathcal{D}_1 = \emptyset$, or $\mathcal{D}_1 = [0, 1]^d$. Therefore, since $[0, 1]^d$ is connected, the RGG is percolating if $r > r_p$, and not percolating if $r < r_p$, where

$$r_p = \frac{1}{(\Omega_d N)^{1/d}}, \quad (\text{B2})$$

where Ω_d is the solid angle in d dimension. Despite our crude approximation in neglecting the node fluctuations, the dependence of r_p on N turns out to be correct (see also [15]). In fact, in $d = 2$ dimension the percolation threshold r_t , *i.e.*, the minimal value of r where a giant connected component appears, is rigorously known to be $r_t \propto (1/N)^{1/2}$. However, the critical value of r above which the RGG is also connected, is rigorously known to be $r_c \propto (\ln(N)/N)^{1/2} > r_t$, therefore the argument of the above approximation is not fully consistent. However, it has the advantage that it is general and can be applied to any sprinkle $\rho(h)$, not necessarily uniform or nearly uniform. Clearly, the smaller is the variance of $\rho(x)$, the better is the approximation.

-
- [1] M. Mezard, G. Parisi, M.A. Virasoro, 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific).
 - [2] O. C. Martin, R. Monasson, R. Zecchina, "Statistical mechanics methods and phase transitions in optimization problems", *Theoretical Computer Science* **265** (2001).
 - [3] L. A. Zdeborová, "Statistical Physics of Hard Optimization Problems", *Acta Physica Slovaca* **59**, 169 (2009).
 - [4] R. Albert, A.L. Barabasi, *Rev. Mod. Phys.* **74** 47 (2002); S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks* (Oxford University Press: Oxford, U.K., 2003); M. E. J. Newman, *SIAM Review* **45**, 167 (2003).
 - [5] G. Bianconi and M. Marsili, *Europhys. Lett.* **74**, 740746 (2006).
 - [6] E. Estrada, N. Hatano, *Chem. Phys. Lett.* **439** 247 (2007).
 - [7] G. Bianconi, *Eurphys. Lett. E* **81**, 28005 (2008).
 - [8] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, *Rev. Mod. Phys.* **80**, 1275 (2008).
 - [9] S. N. Dorogovtsev, *Lectures on Complex Networks* (Oxford Master Series in Statistical, Computational, and Theoretical Physics, 2010).
 - [10] G. Caldarelli, G. A. Capocci, P. De Los Rios, M. A. Muñoz, *Phys. Rev. Lett.* **89**, 258702 (2002);
 - [11] M. Boguñá and R. Pastor-Satorras, *Phys. Rev. E* **68**, 036112 (2003).
 - [12] J. Park and M. E. J. Newman, *Phys. Rev. E* **70**, 066146 (2004).
 - [13] M. Catanzaro and R. Pastor-Satorras, *Eur. Phys. J. B.* **44**, 241 (2005).
 - [14] R. Meester, and R. Roy, "Continuum percolation" (Cambridge University) (1996).
 - [15] J. Dall and M. Christensen, *Phys. Rev. E* **66**, 016121 (2002).
 - [16] M. Penrose, "Random Geometric Graphs" (Oxford Studies in Probability, 5) (2003).
 - [17] P. Kasteleyn, C. Fortuin, *J. Phys. Soc. Japan Suppl.* **26** 1114 (1969).
 - [18] B. Renu, M. Hardwari lal, T. Pranavi, "Routing Protocols in Mobile Ad-Hoc Network: A Review" Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, **115**, 52 (2013).
 - [19] Z. Toroczkai, B. Kozma, K. E. Bassler, N. W. Hengartner and G. Korniss, *J. Phys. A: Math. Theor.* **41**, 155103 (2008).
 - [20] M. Ostilli, *Eurphys. Lett. E* **105**, 28005 (2014).
 - [21] M. Ostilli, *Phys. Rev. E* **89**, 022807 (2014).
 - [22] K. Anand, D. Krioukov, G. Bianconi, *Phys. Rev. E* **89**, 062807 (2014).
 - [23] J. Park and M. E. J. Newman, *Phys Rev E* **70**, 066117 (2004).
 - [24] P. Colomer-de-Simón, M. Á. Serrano, M. G. Beiró, J. I. Alvarez-Hamelin, M. Boguñá, *Sci. Rep.* **3**, 2517 (2013).
 - [25] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá, *Phys. Rev. E* **82**, 036106 (2010).
 - [26] C. I. Del Genio, T. Gross, and K. E. Bassler, *Phys. Rev. Lett.* **107**, 178701 (2011).